Journal of Biomedical Informatics 43 (2010) 587–594

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text

Dina Demner-Fushman *, James G. Mork, Sonya E. Shooshan, Alan R. Aronson

Lister Hill National Center for Biomedical Communications (LHNCBC), U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

ARTICLE INFO

Article history:
Received 9 August 2009
Available online 10 February 2010

Keywords:
UMLS
Metathesaurus
Content views
Natural Language Processing
Indexing
Clinical text

ABSTRACT

Identification of medical terms in free text is a first step in such Natural Language Processing (NLP) tasks as automatic indexing of biomedical literature and extraction of patients' problem lists from the text of clinical notes. Many tools developed to perform these tasks use biomedical knowledge encoded in the Unified Medical Language System (UMLS) Metathesaurus. We continue our exploration of automatic approaches to creation of subsets (UMLS content views) which can support NLP processing of either the biomedical literature or clinical text. We found that suppression of highly ambiguous terms in the conservative AutoFilter content view can partially replace manual filtering for literature applications, and suppression of two character mappings in the same content view achieves 89.5% precision at 78.6% recall for clinical applications.

Published by Elsevier Inc.

1. Introduction

The semantic analysis of biomedical text and mediation between the language of users accessing the biomedical documents for various purposes and the language of the documents depend strongly on the formal representation of the domain language and knowledge [1]. The Unified Medical Language System® (UMLS®) represents, in machine-readable form, information about the biomedical language and domain knowledge and serves as foundation for biomedical language processing. Several questions naturally occur due to the availability of such global knowledge and language resources: (1) how suitable is the resource for a specific goal in terms of coverage? (2) what is the most effective approach to use the resource? (3) can the resource be automatically customized? and (4) are the same customization methods applicable for different tasks, subdomains, and sublanguages?

The suitability of the UMLS for construction of a lexicon for automatic processing of medical narrative was studied by Johnson using the 1997 UMLS SPECIALIST Lexicon and Metathesaurus® [2]. In this study, the SPECIALIST Lexicon covered about 79% of syntactic information and 38% of semantic information in discharge summaries. When the same methodology was applied to construction of a lexicon for processing texts in the field of molecular biology, over 77% of the tokens in the domain corpus were found in the derived lexicon, but only 3% of the unique tokens in the corpus were covered [3]. The UMLS was found to cover approximately 92% of unique concepts in answers to translation research questions (excluding questions about mutations) [4]. In an evaluation of the UMLS as a source of knowledge for processing of chest X-ray reports and discharge summaries, the UMLS-based lexicon did not perform as well as the custom built lexicon that contained most clinical terms found in reports associated with these domains [5]. The authors, however, found the UMLS to be a valuable resource for medical language processing because it substantially reduced the effort in construction of the lexicon. UMLS customization through intersection with local vocabularies was further explored in a study that included lexicons submitted by seven large scale healthcare institutions and resulted in creation of the CORE (Clinical Observations Recording and Encoding) Subset of SNOMED CT® (Systematized Nomenclature of Medicine – Clinical Terms®).[1]

The UMLS Metathesaurus, the major component of the UMLS, is constructed from over 100 biomedical vocabularies. Terms from different vocabularies meaning the same thing are grouped together into concepts, and each concept is assigned one or more categories, or semantic types, from the UMLS Semantic Network. This organization of biomedical concepts consisting of surface forms from UMLS constituent vocabularies serves as a powerful basis for supporting biomedical applications as shown by many studies including those cited above [2–5]. However, Metathesaurus content is known to have a number of problems such as missing

* Corresponding author. Address: Staff Scientist, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bldg. 38A, Room 10S-1020, 8600 Rockville Pike MSC-3825, Bethesda, MD 20894, USA. Fax: +1 301 402 0341.

E-mail address: ddemner@mail.nih.gov (D. Demner-Fushman).

[1] http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html.

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2010.02.005

biomedical concepts, concepts that are not biomedical at all, and ambiguity, perhaps the most important problem with Metathesaurus content.

One source of Metathesaurus ambiguity arises when a concept contains a term which is a substring of the preferred name of the concept but whose meaning differs from that of the concept. For example, the concept *Other location of complaint* contains the term *Other*, which is a spurious synonym of *Other location of complaint*. Note that the Metathesaurus editors have marked some problematic terms as *suppressible*, making them easy to ignore. Unfortunately, the term *Other* in the above example is not so marked. A source of true lexical ambiguity arises from the existence of acronym/abbreviation terms. For example, the term *PAP* occurs in 15 concepts including *Papaverine*, *PAPOLA gene* and *PULMONARY ALVEOLAR PROTEINOSIS, ACQUIRED*. As a final example of true ambiguity, consider the term *resistance*, which occurs as a term in the three concepts *Resistance (Psychotherapeutic)*, *resistance mechanism*, and *social resistance*. Each of these concepts can legitimately be represented by the homonym *resistance*. The problem in this case is that at least one more legitimate sense of resistance, namely *Electrical resistance*, is missing.

As the Metathesaurus has grown, the goal of effectively using its knowledge has become more challenging, partly due to the growth in ambiguity described above. A large body of work on disambiguation of Metathesaurus homonyms in context provides a means for selecting the correct concept [6–12]. This paper presents an alternative approach that attempts to reduce the amount of ambiguity and the size of the resource in the hope of subsequently reducing text processing time and complexity without loss in coverage and accuracy. This reduction of excessive and spurious ambiguity could be of help on its own or combined with word sense disambiguation programs such as the one based on journal descriptor indexing [12] which is an optional feature available in the current MetaMap processing [13].

MetaMap, a tool that identifies Metathesaurus concepts in free text, was used as an essential part of text processing in all experiments presented in this paper. MetaMap employs two data models, relaxed and strict, that differ in how much Metathesaurus content is filtered out [14]. The relaxed model filters out lexically similar strings based on case and hyphen variation, possessives, comma uninversion, *NOS* variation, and non-essential parentheticals. It also includes the manual removal of some strings such as numbers, single alphabetics, *NEC* terms, Enzyme Commission (EC) terms, the short forms of brand names and, most importantly, unnecessarily ambiguous terms [15]. MetaMap's strict model also filters out strings with complex syntactic structure; these are strings which MetaMap does not match well anyway. Table 1 presents examples of removed strings. Over 40% of Metathesaurus strings are removed in the creation of the strict model. It is MetaMap's default model for semantic NLP processing, and it has been available as the first Metathesaurus *Content View* since the 2005AA UMLS release [16].

The Lister Hill NLP Content View (LNCV) project was launched in 2007 to study the effective use of the Metathesaurus and answer questions about applicability of automatic customization methods for different sublanguages [17]. We automatically constructed several Metathesaurus subsets, (called content views) that we hoped might improve the performance of two NLP applications: the NLM Medical Text Indexer (MTI) [18], a literature application, and the Clinical Question Answering [19] clinical application.

Table 1
String filtering in the MetaMap strict model.

Metathesaurus strings	Reason(s) for removal
Intraductal carcinoma, non-infiltrating NOS (morphologic abnormality)	*NOS* variation, comma uninversion, parenthetical, case, hyphen
[D] Castleman's disease (disorder)	Parenthetical, case, possessive
[M] Hodgkin's sarcoma	Parenthetical, case, possessive
[X] Diffuse non-Hodgkin's lymphoma, unspecified (disorder)	Comma uninversion, parenthetical, case, hyphen, possessive

1.1. Applications used to evaluate the Metathesaurus content views

1.1.1. Medical Text Indexer (MTI)

MTI is a system for producing indexing terms, either Medical Subject Headings (MeSH®) or Heading/Subheading combinations, from biomedical text. It has been used at NLM since 2002 in both semi-automated and fully automatic environments. MTI indexing recommendations are available to NLM indexers to assist them in indexing MEDLINE® citations, and the recommendations are consulted for about 40% of MEDLINE indexing. MTI also assists NLM catalogers and the History of Medicine division, and produces fully automatic indexing (subject to selective review) for collections of abstracts available through the NLM Gateway [20].

1.1.2. Clinical Question Answering (CQA)

The CQA Clinical Question Answering system represents questions and MEDLINE citations using frames which capture the fundamental elements of Evidence Based Medicine (EBM) [21]: (1) clinical scenario; (2) clinical task (diagnosis, therapy and prevention, prognosis, and etiology); and (3) strength of evidence [19]. Given a clinical note, the system automatically generates a question frame using MetaMap [13] and a set of rules for extraction of the elements of a clinical scenario. The question frame is used to automatically generate a query and search MEDLINE. Retrieved citations are processed with several knowledge extractors and classifiers that rely on a combination of factors: UMLS concept recognition using MetaMap, manually derived patterns and rules, and supervised machine learning techniques to identify the fundamental EBM components. The answers in the form of patient oriented outcome statements are extracted from retrieval results retained after fuzzy unification of the question and answer frames [19].

To generate question frames, the CQA system extracts from the MetaMap output concepts that belong to the following semantic groups: *Problems/findings* (meant to represent a patient's problem list), *Interventions*, and *Anatomy* (which provides details about the patient). The *Problems/findings* semantic group is based on the UMLS semantic group *Disorders* [22] augmented with semantic types *Laboratory or Test Result*, *Virus* and *Bacterium* because in clinical narrative, entities of those types could be treated as findings. For example, the phrase "*Urine Cx results + for non-fermenter not pseudomonas*" means the patient tests positive for non-fermenting bacteria. The *Interventions* group includes therapeutic and diagnostic procedures, drugs, and drug delivery devices. The *Anatomy* group includes semantic types in the anatomy and physiology groups excluding those on the cell and molecular level (for example, *Cell* or *Molecular Function*).

1.2. Previously evaluated Metathesaurus content views

In our 2008 study, we designed experiments to determine if any of the content views could improve the performance for either the literature (MTI indexing) or clinical application (extraction of the answer frames) [17]. In the 2008 study, we took two disparate approaches for defining the content views, a maximalist approach (which strives to maximally retain the Metathesaurus strings) and a minimalist approach (which reduces the Metathesaurus size to a minimum).

The maximalist approach, patterned after the data model construction used by MetaMap [13], consists of progressive removal of Metathesaurus strings when they are determined to be inappro-

priate for the data model being built. In contrast, the minimalist approach begins by removing a significant portion of the Metathesaurus to form a minimal set and then restoring useful strings in a backoff phase. In our case, we removed all concept strings that were a proper substring (respecting word boundaries) of another string in the concept to form the minimal set. Examples of these and other content view modifications are provided in Section 2.

We identified three maximal content views (Base, AutoFilter, and AllFilter) that performed well for the literature application (MTI) and two minimalist content views (MinBackoff and Minimal)[2] that performed well for the clinical application (CQA). Of these, we chose the three best performing content views (AutoFilter, AllFilter, and Minimal) for further study.

1.3. Modifications to previously evaluated Metathesaurus content views

For the current study, we have developed three content view modification approaches – *conservative*, *moderate*, and *aggressive* – designed to systematically remove more and more Metathesaurus strings from a content view as they progress from *conservative* to *aggressive*. We used the same NLP applications (MTI and CQA) and supplemented our 2008 LNCV document collection with a collection of randomly selected de-identified clinical discharge sentences to evaluate the extraction of question frames.

The three major approaches were designed to expand on our earlier LNCV work. The *conservative* approach deleted some short Metathesaurus strings that we thought were contributing to the overall ambiguity. The *moderate* approach removed specific source vocabularies that, in the analysis preceding this work, were shown to rarely be a single source for terms found in MEDLINE and were introducing possibly ambiguous and/or incomplete concept senses. The *aggressive* approach performed a wholesale removal of blocks of Metathesaurus strings based on their degree of ambiguity.

2. Methods

2.1. Experimental environment

As mentioned earlier, we are reusing three of last year's five content views extracted from the 2007AB Metathesaurus (English strings only) – AutoFilter, AllFilter, and Minimal. The AutoFilter view consists of all of MetaMap's automatic filtering; it is MetaMap's strict model but without the manual ambiguity filtering. The AllFilter view is MetaMap's strict model, including manual filtering. The Minimal view, the most restrictive of all the content views, removes all Metathesaurus strings that are a proper substring of another string in the same concept, respecting word boundaries. For example, the string *malaria* is removed from the concept *Malaria Vaccines*. The order of the Metathesaurus content views from most conservative to most aggressive is AutoFilter, AllFilter, and Minimal.

The *conservative* and *moderate* content view modification approaches are based on results of our manual ambiguity review process that is performed each year on the "AA" Metathesaurus release. The goal here is to automate some of the manual processes to improve performance and to allow us to do deeper manual review of the remaining ambiguities. We chose the source vocabularies for the *moderate* approach partially based on data from an unpublished internal study that showed how much of each of the UMLS source vocabulary strings were actually found in MEDLINE®/PubMed®.

[2] Formerly called AggrBackoff and Aggressive.

Three *conservative* approaches consisted of (1) removing all Metathesaurus strings with 2 characters (for example, *ds* where MetaMap returns *diethyl sulfate*, *DHDDS gene*, *DHPS wt Allele*, *DS*, *Disposition Submission Domain*, and *Supernumerary maxillary right lateral primary incisor*); (2) all strings with 3 characters (for example, *not* where MetaMap returns *NR4A2 gene*, and *Negation*; and (3) all 3-character consonants (for example, *pcr* where MetaMap returns *Polymerase Chain Reaction*). The effects of deletion of each type of short strings were studied separately.

The *moderate* approach involved the complete removal of specific source vocabularies that we thought were either contributing possibly ambiguous and/or incomplete senses to our MetaMap results or which contained large numbers of terms not likely to appear in biomedical text. The vocabularies we removed were HL7 (Health Level Seven), LOINC (Logical Observation Identifier Names and Codes Vocabulary), and RXNORM (RxNorm Vocabulary). HL7 is an example of a vocabulary which, when added to the Metathesaurus, contributes many ambiguous terms and incomplete senses. Also, we discovered in an internal study that LOINC and RXNORM are examples of large Metathesaurus vocabularies whose terms occur less than 5% of the time in MEDLINE, a good source of biomedical text.

The *aggressive* approach consisted of the removal of blocks of Metathesaurus strings based on their degree of ambiguity. For these experiments we concentrated on 2+ ambiguities through 10+ ambiguities. The plus sign after the number indicates that strings with the given degree of ambiguity and higher were removed. Degree of ambiguity is based on the number of senses for a given UMLS concept after all of the UMLS identified suppressible senses are removed via part of the MetaMap data creation process. In the UMLS MRCON file, senses are marked as suppressible by the lowercase "s" in the third column. For example, *abdomen* is five ways ambiguous in the Metathesaurus, but two of the senses are already marked as suppressible in the UMLS leaving MetaMap with *abdomen* being a three ways ambiguous concept.

Fig. 1 summarizes the content views, modifications applied to each view, and document collections used in our experiments.

2.2. LNCV and clinical document collections

The set of documents used in MTI experiments, the 2008 LNCV document collection, consists of a randomly chosen subset of 10,000 MEDLINE citations indexed in 2007 that had MTI recommendations. The clinical text collection consists of 356 random sentences, each from a different randomly selected de-identified discharge summary obtained from the Laboratory for Computational Physiology, Massachusetts Institute of Technology [23,24].

2.3. Experiments

We repeated the 2008 baseline experiments [17] for the AllFilter, AutoFilter, and Minimal content views to verify that the new 2009 results were consistent with the original 2008 results. While the tools themselves did not change, MTI specifically relies on the related citations algorithm applied to the ever changing PubMed database for part of its data. Related citations did contribute to an insignificant difference in results which we then used as our new baselines for this round of experiments.

All 16 content view modification experiments (three short string, four UMLS source vocabulary, and nine ambiguity) were run on all three content views (AutoFilter, AllFilter, and Minimal) for a total of 48 experiments on this new baseline, producing results for both document collections.

Both the literature and clinical applications used the following criterion for the *conservative* and *aggressive* experiments: concepts that would have been found only using a string removed from the

Content views

- AutoFilter: MetaMap's strict model without the manual ambiguity filtering
- AllFilter: MetaMap's strict model with the manual ambiguity filtering
- Minimal: substrings suppressed

Content view modifications

- Short string removal (*conservative*): UMLS concepts of 2 characters, 3 characters, and 3 character consonants
- UMLS source vocabulary removal (*moderate*): HL7, RXNORM, LNC, and all three combined
- Ambiguity removal (*aggressive*): 2+ through 10+ ambiguity

Document collections

- 2008 LNCV document collection: 10,000 MEDLINE citations
- Clinical text collection: 356 random de-identified discharge sentences

Fig. 1. Data summary.

content view were not mapped. However, if a MeSH Heading (for MTI) or UMLS concept (for CQA) could have been reached by more than one triggering string and one of those triggers was not removed, we kept the MeSH Heading/UMLS concept. For example, if we were removing all the UMLS strings of three characters from the MTI results and we had MeSH Heading *Immunoglobulin G* triggered by two strings *IgG* and *Immunoglobulin G* found in the same MEDLINE abstract, we would keep *Immunoglobulin G* in this case because it was triggered by the longer string *Immunoglobulin G*. Conversely, mapping the string *ds* in the clinical note to *Supernumerary maxillary right lateral primary incisor* was removed because it was found only through the two character string in the note.

2.3.1. Literature (MTI) experiments

The MTI experiments consisted of processing the 2008 LNCV document collection through MTI [18] using one of the content views and one content view modification criterion defined in Fig. 1 above. Since the *moderate* experiments entailed the exclusion of one of three UMLS source vocabularies, they were conducted by replacing the normal MetaMap data model used by MTI with one of three data models constructed after removing one of the vocabularies from the Metathesaurus. Performing the *conservative* and *aggressive* experiments was simpler: baseline MTI results using MetaMap's normal data model were modified by removing those MeSH Headings meeting the specific criteria for the experiment as described above.

2.3.1.1. MTI indexing evaluation. The indexing recommendations so obtained were compared with the official MeSH indexing for the documents, computing recall (R), precision (P), and F_2 values for each document. The *F*-measure $F_2 = 5 * (PR)/(4P + R)$ gives recall twice as much weight as precision in order to reflect the indexing perspective that finding additional relevant indexing terms is more important than including a few irrelevant terms.

2.3.2. Clinical (CQA) experiments

The CQA *moderate* experiments involved processing the clinical text collection through MetaMap replacing the normal MetaMap data model with a data model constructed after removing the source vocabularies from the Metathesaurus and then run against *Problems*, *Interventions*, and *Anatomy* extraction facilities. The *Problems* and *Interventions* extractors identify two of the four elements of a well-formed clinical question frame [25]. The *Anatomy* extractor contributes to the *Patient/Problem* element of the frame. The complete removal of vocabularies in *moderate* content view modifications requires a specific data model to be used in MetaMap processing. Conversely, the *conservative* and *aggressive* modifications allow post-processing the results obtained using the normal MetaMap data model. For the *conservative* and *aggressive* experiments, the baseline CQA results for each content view were used, and UMLS concepts were removed from the baseline CQA results when the specific criteria were met as described above.

2.3.2.1. CQA extraction evaluation. To identify the most suitable UMLS customization approach, the reference standard for the clinical application was created as follows: *Problems/findings*, *Interventions*, and *Anatomy* terms were manually annotated by DDF prior to the evaluation. The evaluation of the modifications to the UMLS content views was conducted manually by DDF who matched the UMLS concepts extracted by the system from each sentence into its question frame to the reference standard. The evaluation was conducted manually because we did not see a good way to automate the semantic (rather than lexical) matching process. For example, *temp* in *Temp 97.1* was annotated as shorthand for *temperature measurement (Intervention)* in the reference standard. The term *temp* was also identified as *Intervention (therapeutic procedure)* by MetaMap. However, the preferred name for the concept (and its surface representation *temp*) identified by MetaMap is *cisplatin/etoposide/mitoxantrone/tamoxifen protocol*, which clearly indicates a false positive mapping. To avoid counting such occurrences as true positives, each automatically extracted term was manually compared to the previously created reference standard and evaluated as true positive, false positive, or false negative.

Figs. 2 and 3 present examples of extracted sentences, annotated reference frames, and frames generated by the system. No annotation beyond entity recognition (for example, negation, temporal relations or the severity of problems) was undertaken. The frames generated by the system (in column 3) were compared to the reference frames (column 2). The system frames contain the input string that was matched, the preferred UMLS name for the concept to which the string was matched and its semantic type. The system output is shown for the baseline Minimal view. The conservative modification of this view suppresses the three false positive concepts in the second example (*dm*, *os*, and *hs*) shown in Fig. 3 (as highlighted by the strikethrough text). However, suppression of mappings triggered by three character strings often leads to loss in recall (as shown in Fig. 2 by the strikethrough text).

The discharge summary sentences contained 928 named entities (505 *Problems/findings*, 314 *Interventions*, and 109 *Anatomy* terms). The entities were annotated "as is": without any changes to the original text (such as spelling corrections or abbreviation expansions.) For example, in *Broad spectrum abx started*, "*abx*" (*antibiotics*) was annotated as an intervention. Recall, precision,

an MRI of the sacral area near the abscess ruled out osteomyelitis.		
	Reference annotation	**System output**
Problem(s)	abscess osteomyelitis	Abscess(Abscesses)[patf] osteomyelitis(Osteomyelitides)[dsyn]
Intervention(s)	MRI	~~mri(Magnetic resonance imaging)[diap]~~
Anatomy	sacral area	

Fig. 2. Example sentence extracted from discharge summaries and annotated for the evaluation. Strikethrough indicates mappings suppressed in conservative modification triggered by three character strings (for example, *mri*).

The patient is a 43 year old male with type 1 DM treated with insulin pump, complicated by gastroparesis, mild-moderate retinopathy and neuropathy with several recent admissions for DKA presents with 1 day h/o nausea and vomiting.		
	Reference annotation	**System output**
Problem(s)	type 1 DM gastroparesis retinopathy neuropathy DKA nausea vomiting	 gastroparesis(Gastropareses)[dsyn] mild retinopathy(Mild retinopathy)[fndg] neuropathy(NEUROPATHY)[dsyn] vomiting(Vomiting)[sosy]
Intervention(s)	insulin pump	~~dm(Hexadecadrol)[phsu, strd]~~ insulin pump(Insulin Pumps)[medd]
Anatomy	male	male(Male gender)[orga] ~~os(Skeletal bone)[bpoc]~~ ~~hs(Supernumerary maxillary left primary canine)[bpoc]~~

Fig. 3. Example sentence extracted from discharge summaries and annotated for the evaluation Strikethrough indicates mappings suppressed in conservative modification triggered by two character strings.

and F_2 values for each semantic group and for all entities were computed.

2.3.3. Evaluating experimental results for statistical significance

We used a two-tail paired *t*-test to determine if the differences observed in the MTI and clinical entity extraction experiments are statistically significant. For MTI experiments we are only interested in the significant differences in the F_2 scores as the indicator of the overall improvement in the system. For the clinical application, we are interested in differences in all metrics because in some situations recall is more important than precision (for example, for a clinical researcher in an exploratory task), whereas in some other clinical tasks (for example, retrieving literature to provide clinical evidence) precision is more important than recall.

3. Results

Tables 2 and 3 present the best results compared to the baseline experiments for both applications. (Note that in all these experiments we evaluate not the performance of the tools, but rather use the differences in the performance of the tools to evaluate the approaches to UMLS customization.) The results for the remaining experiments did not improve the baseline significantly, or performed worse. These results are available in the appendices. The bold text in both Tables 2 and 3 indicates the best performing experiments. Table 2 presents the best MTI results together with their baselines. All of the best MTI results involve an aggressive content view modification consisting of removal of ambiguities of a certain degree or higher. The differences in *aggressive modifications* of the three content views compared to the baseline performance of the views are statistically significant ($p < 0.001$). The *aggressive modification* of the Minimal view is significantly worse than the baseline AllFilter view. The difference between the *aggressive modification* of the AutoFilter view and the baseline AllFilter view is not statistically significant. It is important to compare the results to the AllFilter baseline because it is the currently available MetaMap model which we hope to improve.

The results of the *aggressive modification* approach indicate that we might be able to automate some of the ambiguity study that we now do manually. The table includes descriptive information at the beginning as well as three sections of results: the overall results, title-only citations, and those with both title and abstract.

Table 3 contains the results for *conservative* modifications to the three UMLS content views (AllFilter, AutoFilter, and Minimal) evaluated in CQA extraction experiments. Asterisks (∗) indicate statistically significant differences in the overall results for 2- and 3-character term elimination within the same content view. Section signs (§) indicate significant differences between the experimental views and the currently available MetaMap view (AllFilter). Bold typeface indicates the highest recall, precision, and *F*-score for each extractor and for the extraction task overall.

The suppression of whole vocabularies (LOINC, RXNORM, and HL7) did not change the extraction results significantly. The same is true for the terms with high ambiguity. There were no terms with eight or more senses in the discharge sentences and only one term with seven senses, *CAD*, which occurred in four sentences. These four instances of *CAD* contributed to false negatives and false positives in the Minimal and AllFilter content views and to true positives (*Problem* sense) and false positives (*Intervention* sense) in the AutoFilter content view, but not sufficiently to change the results. This term (*CAD*) was also most frequent within five and six or more senses, which led to results similar to 7+ ambiguity suppression. Removal of the terms with two, three, and four

Table 2
MTI results for all content views with the best *aggressive* experiments.

	AutoFilter baseline	AutoFilter 7+ ambig	AutoFilter baseline	AutoFilter 7+ ambig	Minimal baseline	Minimal 6+ ambig
Citations	9999	9999	9999	9999	9999	9995
Indexed MHs	115,877	115,877	115,877	115,877	115,877	115,877
MTI recommendations	187,721	186,701	187,186	186,748	180,113	179,694
All citations						
Correct MTI recommendations	58,417	58,384	58,464	58,443	56,174	56,155
% of Indexed MHs (recall)	50.41%	50.39%	50.45%	50.44%	48.49%	48.47%
% of MTI recommendations (precision)	31.12%	31.27%	31.23%	31.30%	31.19%	31.25%
F_2	44.85%	**44.90%**	44.92%	**44.94%**	43.65%	43.66%
Title-only citations						
Correct MTI recommendations	2745	2740	2753	2748	2562	2561
% of Indexed MHs (recall)	19.81%	19.78%	19.87%	19.83%	18.51%	18.51%
% of MTI recommendations (precision)	44.35%	44.87%	44.74%	44.93%	44.22%	44.70%
F_2	**22.28%**	**22.27%**	22.36%	22.32%	20.95%	20.97%
Title/abstract citations						
Correct MTI recommendations	55,672	55,644	55,711	55,696	53,612	53,594
% of Indexed MHs (recall)	54.57%	54.54%	54.61%	54.59%	52.55%	52.53%
% of MTI recommendations (precision)	30.67%	30.81%	30.77%	30.83%	30.76%	30.81%
F_2	47.21%	**47.26%**	47.28%	**47.30%**	46.03%	46.04%

Table 3
CQA extraction results for all content views with *conservative* 2- and 3-character string suppression experiments for each semantic group and overall.

	AutoFilter baseline	AutoFilter 2 char	AutoFilter 3 char	AllFilter baseline	AllFilter 2 char	AllFilter 3 char	Minimal baseline	Minimal 2 char	Minimal 3 char
Problems									
Recall	**79.80%**	79.21%	75.24%*	77.82%	77.82%	74.06%	67.52%	67.13%	66.34%
Precision	90.15%	**93.67%**	90.90%	90.97%	92.91%	92.57%	83.57%	92.87%	86.12%
F_2	81.68%	**81.73%**	77.92%*	80.14%	80.43%	77.15%	70.22%	71.07%	69.53%
Interventions									
Recall	**79.94%**	77.39%	76.11%*	76.11%	76.11%	73.24%	60.19%	59.23%	58.59%
Precision	77.71%	83.21%	82.99%	81.84%	**83.56%**	88.46%	71.59%	77.82%	80.70%
F_2	**79.48%**	78.49%	77.39%*	77.19%	77.49%	75.85%	62.17%	62.20%	61.99%
Anatomy									
Recall	**79.82%**	**79.82%**	77.06%*	79.81%	79.81%	77.06%	76.14%	76.14%	72.47%
Precision	74.36%	90.63%*	76.36%*	88.78%	92.55%	92.31%	70.33%	**96.51%**	70.53%
F_2	78.66%	81.77%*	76.92%*	81.46%	**82.07%**	79.69%	74.90%	79.50%	72.07%
Overall									
Recall	**79.85%**§	78.66%	75.75%*	77.48%	77.48%	74.14%*	66.06%§	65.52%	64.44%
Precision	83.54%	89.57%*	86.15%*	87.47%	89.54%	**91.13%**	77.59%§	88.12%*	82.03%
F_2	80.56%	**80.62%**	77.62%*	79.29%	79.62%	77.01%	68.08%§	69.06%*	67.33%

senses from the best overall performing model, AutoFilter, degraded the results (see Table 4).

4. Discussion

Before discussing the results for the MTI and problem and intervention extraction experiments individually, it is worth observing that the MTI experiments scored far lower for all measures than the extraction experiments. In general this is due to the fact that MTI's indexing task is more complex and challenging than the extraction task. Specifically, MEDLINE indexing involves the creation of a list of about 12 interrelated terms, which together characterize the essence of a biomedical article. Furthermore, MTI produces up to 25 indexing terms for a given article and is therefore penalized when its indexing recommendations are compared via exact matching with the more parsimonious MEDLINE indexing.

Table 4
Best AutoFilter extraction results were degraded by suppression of low ambiguity terms.

	Ambiguity 4+	Ambiguity 4+ 2 char	Ambiguity 4+ 3 char	Ambiguity 3+	Ambiguity 3+ 2 char	Ambiguity 3+ 3 char	Ambiguity 2+	Ambiguity 2+ 2 char	Ambiguity 2+ 3 char
Recall	77.58%	76.40%	73.92%	75.75%	74.56%	72.27%	68.17%	67.03%	64.55%
Precision	84.61%	89.86%	87.06%	84.69%	90.10%	87.13%	84.65%	89.88%	86.94%
F_2	78.89%	78.76%	76.22%	77.38%	77.22%	74.82%	70.93%	70.62%	68.06%

In general, the MTI results were disappointing for both the *conservative* and *moderate* experiments. MTI did worse for all of the experiments, except for title-only citations in the *moderate* experiments, where it performed slightly better in most cases. This slight improvement in results may be due to title-only citations having a smaller list of recommendations (less than 6 vs. 25 or more for regular citations) and/or our use of Word Sense Disambiguation settings for MetaMap processing. Both of these methods create a smaller more precise list, and our experiments may have removed further problematic recommendations. The positive outcome of our experiments is the observation that 7+ ambiguity removal in AutoFilter view is comparable to the laborious manual review that turns this view into the AllFilter view.

The results of extracting *Problems* and *Interventions* from clinical narrative show not only that processing of the same text for different tasks (MTI and clinical entity extraction) needs different models [17], but also that processing of different text types (citations vs. clinical text) for the same task requires different models. The Minimal view that performed best on MEDLINE citations is significantly worse than other approaches for clinical text processing. Although its precision was improved by the *conservative* and *moderate* modifications (achieving the highest precision, 96.5% for *Anatomy* extraction), its low recall with overall insignificant improvement in precision makes this approach unsuitable for clinical narrative processing. The difference in the validity of this view for extraction of the same entity types from the formal language of MEDLINE citations could be explained by the differences in the two sublanguages. For example, the string *anxiety* is not extracted from the sentence "*Diazepam 2 mg Tablet Sig: One (1) Tablet PO Q8H (every 8 h) as needed for anxiety*." when the mapping is done using the Minimal view. This string maps to three UMLS concepts: *Anxiety Adverse Event* [C1963064]; *Anxiety* [C0003467]; and *Anxiety symptoms* [C0860603]. Whereas *Anxiety Adverse Event* and *Anxiety symptoms* have semantic type *finding* and are found in AllFilter and AutoFilter views as *Problems* through their synonym *anxiety*, in the Minimal view this synonym is suppressed. This mapping is not reached through the concept *Anxiety* [C0003467] because its semantic type is *mental process*, which in most cases does not signify a clinical problem or finding. Identification of *Problems* in the text of MEDLINE citations, in general, is more robust than the same process applied to clinical text. Processing of clinical text mostly depends on a single occurrence of the term, but abstracts of scientific articles repeat the name of the disorder and findings that were studied several times and at least one of those mentions usually provides the full name. For example, 4699 randomized clinical trials (RCT) abstracts in MEDLINE contain terms *anxiety* and *symptom* or *symptoms*, but only 2783 RCT abstracts contain the term *anxiety*, but not *symptom* or *symptoms*.

The AutoFilter and AllFilter views performed equally well (with a significantly higher recall in the AutoFilter view). This indicates that for clinical text the AutoFilter content view with 2-character terms removed can replace the laborious manual filtering process involved in the AllFilter content view.

5. Conclusion

In continuing construction of UMLS content views, we focused on improvement of the three most promising UMLS content views identified in our earlier study [17] Our first goal was to reduce manual effort in constructing the AllFilter content view most suitable for medical text indexing. For MTI, the results obtained using the 7+ *aggressive* modification of the AutoFilter content view are comparable to those achieved by the manually constructed AllFilter content view.

Our second goal was to test if the Minimal content view, most suitable for extraction of the elements of answers to clinical questions from MEDLINE citations, is also appropriate for extraction of the elements of clinical questions from clinical notes. In processing clinical notes, the recall levels for this view were significantly lower than for the other two views, which might be explained by extensive use of abbreviated terms in the clinical notes. The best overall F_2 score was achieved by the AutoFilter content view with 2-character strings suppression.

The fact that we were able to construct fully automatic content views that perform at least as well as manually constructed views is encouraging. Our experiments suggest, however, that content views need to be constructed for each specific task and sublanguage (text type).

The datasets used in these experiments are available through the MetaMap Portal.[3]

Acknowledgments

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine. The authors thank Dr. Roger Mark for permission to make the clinical dataset publicly available.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2010.02.005.

References

[1] McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A, Srinivasan S. UMLS knowledge for biomedical language processing. Bull Med Libr Assoc 1993;81(2):184–94.
[2] Johnson SB. A semantic lexicon for medical language processing. J Am Med Inform Assoc 1999;6(3):205–18.
[3] Verspoor K. Towards a semantic lexicon for biological language processing. Comp Funct Genomics 2005;6(1–2):61–6.
[4] Overby CL, Tarczy-Hornoch P, Demner-Fushman D. The potential for automated question answering in the context of genomic medicine: an assessment of existing resources and properties of answers. BMC Bioinformatics 2009;10(Suppl. 9):S8.
[5] Friedman C, Liu H, Shagina L, Johnson S, Hripcsak G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. Proc AMIA Symp 2001:189–93.
[6] Liu H, Teller V, Friedman C. A multi-aspect comparison study of supervised word sense disambiguation. J Am Med Inform Assoc 2004;11(4):320–31.
[7] Schuemie MJ, Kors JA, Mons B. Word sense disambiguation in the biomedical domain: an overview. J Comput Biol 2005;12(5):554–65.
[8] Xu H, Markatou M, Dimova R, Liu H, Friedman C. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. BMC Bioinformatics 2006;7:334.
[9] Stevenson M, Guo Y, Gaizauskas R, Martinez D. Disambiguation of biomedical text using diverse sources of information. BMC Bioinformatics 2008(9 Suppl. 11):S7.
[10] Savova GK, Coden AR, Sominsky IL, Johnson R, Ogren PV, de Groen PC, et al. Word sense disambiguation across two domains: biomedical literature and clinical notes. J Biomed Inform 2008;41(6):1088–100.
[11] Alexopoulou D, Andreopoulos B, Dietze H, Doms A, Gandon F, Hakenberg J, et al. Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. BMC Bioinformatics 2009;10:28.
[12] Humphrey SM, Rogers WJ, Kilicoglu H, Demner-Fushman D, Rindflesch TC. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: preliminary experiment. J Am Soc Inf Sci Technol 2006;57(1):96–113.
[13] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17–21.
[14] Mork JG, Aronson AR. Filtering the UMLS Metathesaurus for MetaMap: 2009 Edition. National Library of Medicine, Bethesda, MD. Available from: http://skr.nlm.nih.gov/papers/references/filtering09.pdf.
[15] Shooshan SE, Mork JG, Aronson AR. Ambiguity in the UMLS Metathesaurus: 2009 Edition. National Library of Medicine, Bethesda, MD. Available from: http://skr.nlm.nih.gov/papers/references/ambiguity09.pdf.

[3] http://metamap.nlm.nih.gov.

[16] Unified Medical Language System: Preface to the 2005AA Documentation. National Library of Medicine, Bethesda, MD. Available from: http://www.nlm.nih.gov/archive/20080407/research/umls/archive/2005AA/umlsdoc_preface.html.
[17] Aronson AR, Mork JG, Neveol A, Shooshan SE, Demner-Fushman D. Methodology for creating UMLS content views appropriate for biomedical natural language processing. Proc AMIA Symp 2008:21–5.
[18] Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM indexing initiative's medical text indexer. Stud Health Technol Inform 2004;107(Pt. 1):268–72.
[19] Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Comput Linguist 2007;33(1):63–104.
[20] Kingsland 3rd LC, Prettyman MF, Shooshan SE. The NLM Gateway: a metasearch engine for disparate resources. Stud Health Technol Inform 2004;107(Pt 1):52–6.
[21] Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. 2nd ed. Edinburgh, Scotland: Churchill Livingstone; 2000.
[22] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform 2001;84(Pt. 1):216–20.
[23] Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.
[24] Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol 2002;29:641–4.
[25] Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club 1995;123(3):A12–3.